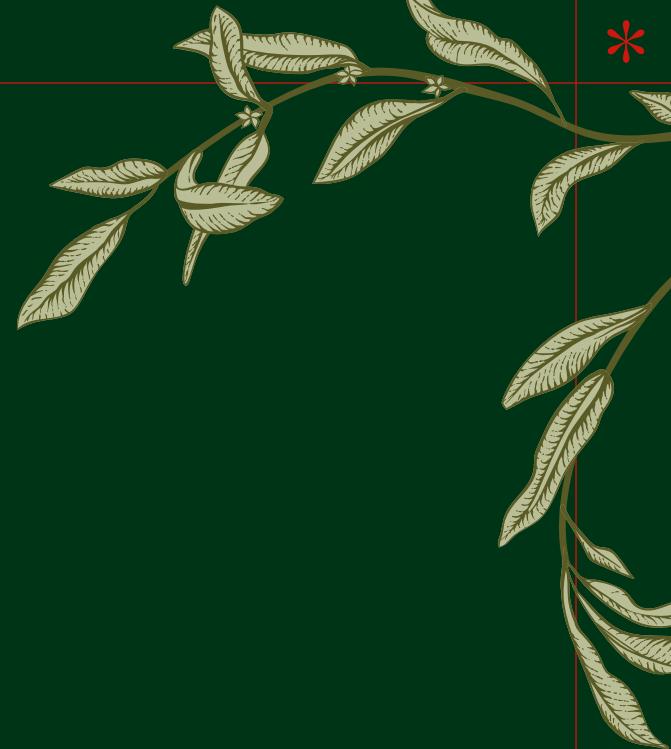


DS 5030 Project 1

A decorative illustration in the bottom left corner features several red-orange flowers with detailed petals and green leaves, arranged in a cluster.

Sophie Kim
Harry Millspaugh
Sheyi Faparusi
Jessica Oseghale
Tiandra Threat
Brooke Lumpkin
Grace George
Caroline Kranefuss



Question 1:

Dataset Overview

- **Source:** EPA Facility-Level Greenhouse Gas Emissions
- **Focus:** *Non-Biogenic* CO₂ emissions
- **Years:** 2011–2023
- **Goal:** Model the distribution of carbon emissions across U.S. states

Data Provenance

- **Collected by:** U.S. Environmental Protection Agency
- **Purpose:** Track facility greenhouse gas emissions for policy & regulation
- **Our Use:** Aggregate facility emissions → state-level totals

Data Quality & Missing Values

- Some incomplete or missing facility/state records
- Large emitters create noticeable outliers
- Cleaning: standardized units, removed nulls, aggregated to state totals

Why This Dataset

- Reliable, government-verified, publicly accessible
- Suitable for modeling how emissions vary across states

Question 2:

Phenomenon

- Modeling year-by-year summed non-biogenic carbon dioxide emissions
 - By state/territory
 - Normalized by area

Background

- Industry emissions account for 23% of total U.S. Greenhouse Gas Emissions in 2022 (1)
- Non-biogenic CO₂
 - Result of non-renewable carbon sources such as coal, oil, natural gas, and petroleum products
 - *Human-caused* CO₂ emissions

Features and support

- Expect larger states/territories to have larger emissions (see Question 3)
- Expect more population-dense states/territories to have larger normalized emissions (see Question 6)

Question 3: ECDF, KDE (Non-Parametric Model)

Steps to get realistic proportions:

1. *Non-Biogenic CO₂ Emissions* selected
 - o *Intriguing*
2. *Aggregate* all states together – one row per State and Year
3. *Emissions per Area* – realistic picture of data

Challenges to Overcome:

1. Large emissions and area
2. Large ranges

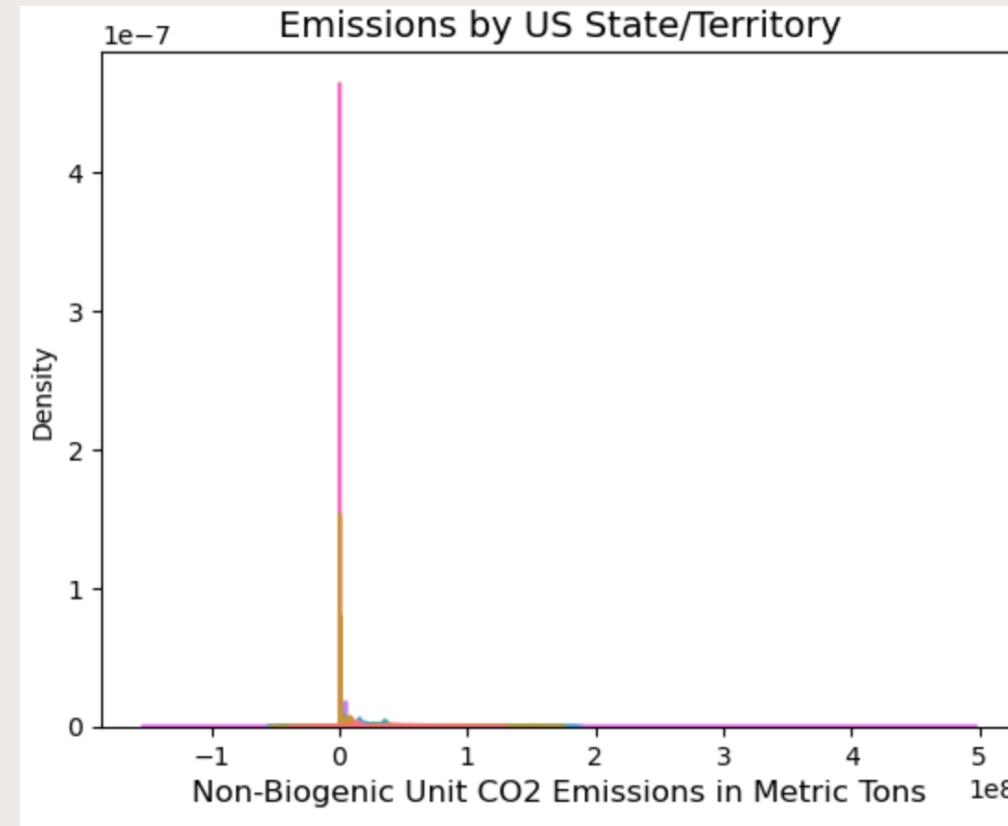
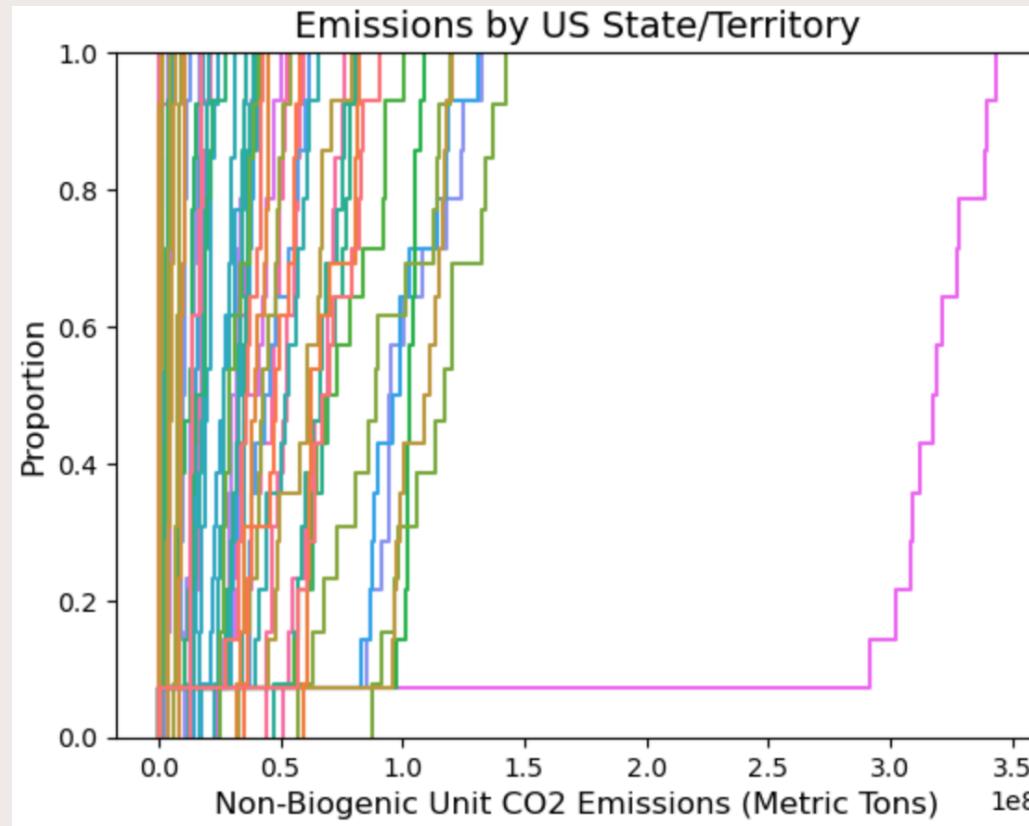
Solutions:

1. Took a *log* – standardize data, easier to visualize
2. Took a *ratio* of emissions to area

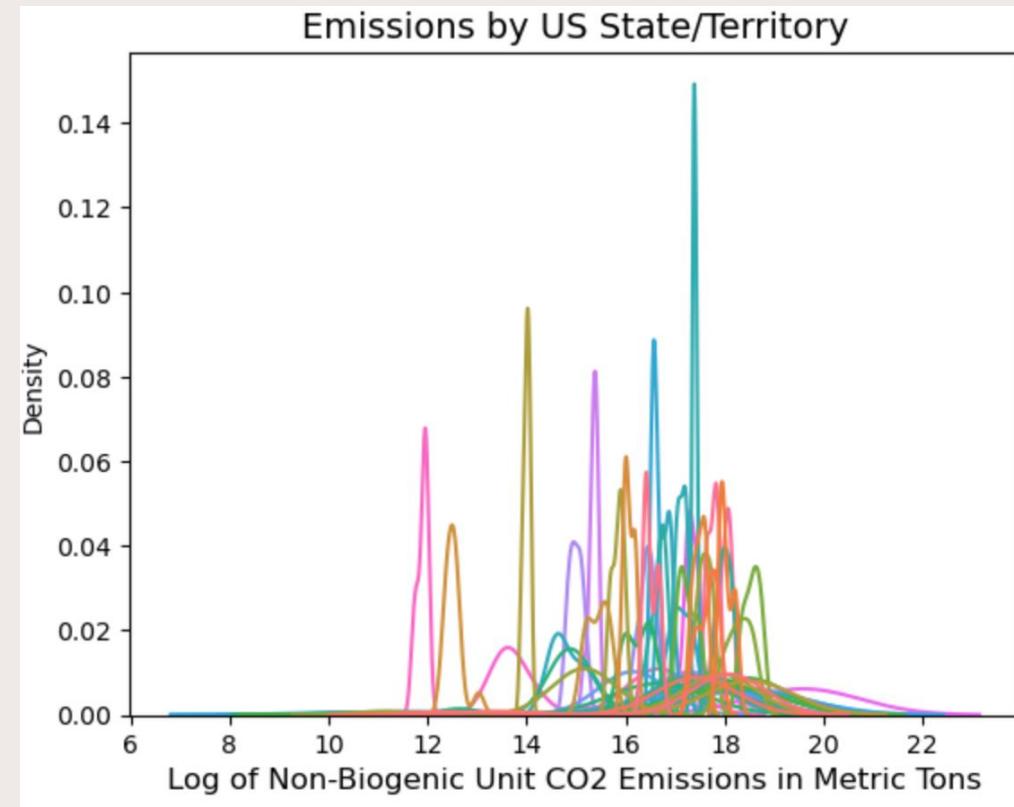
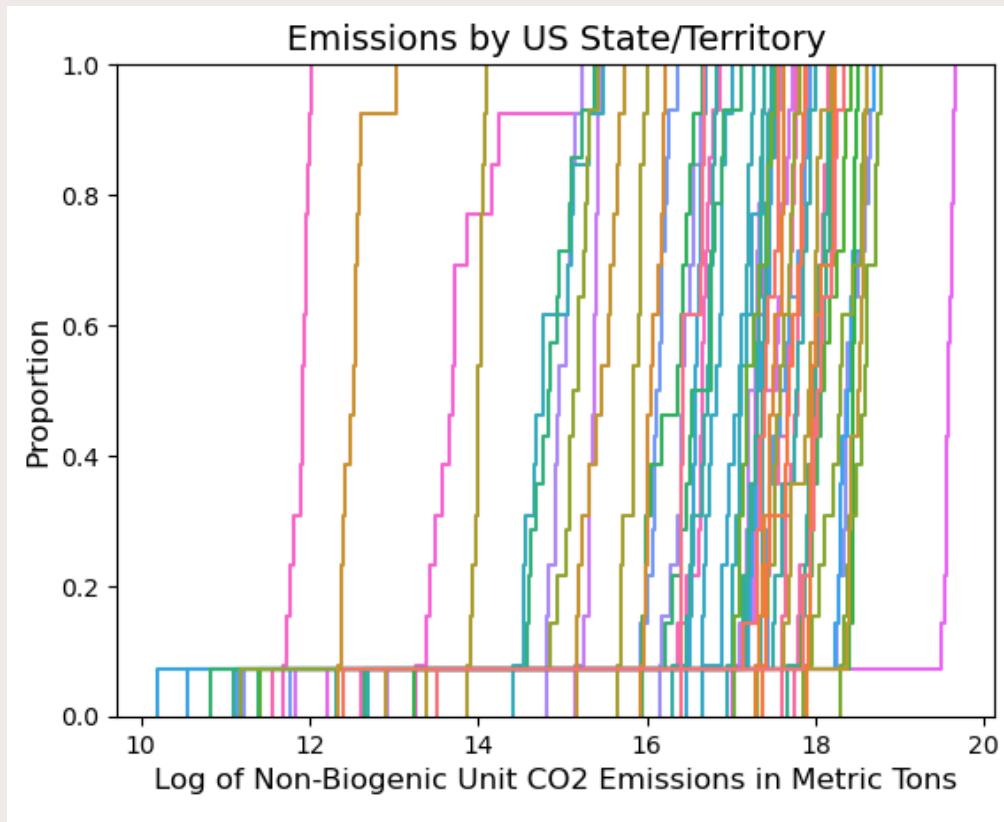
Notes:

1. Overall *quality* of data good
 - o Pre-cleaning

Question 3: ECDF, KDE

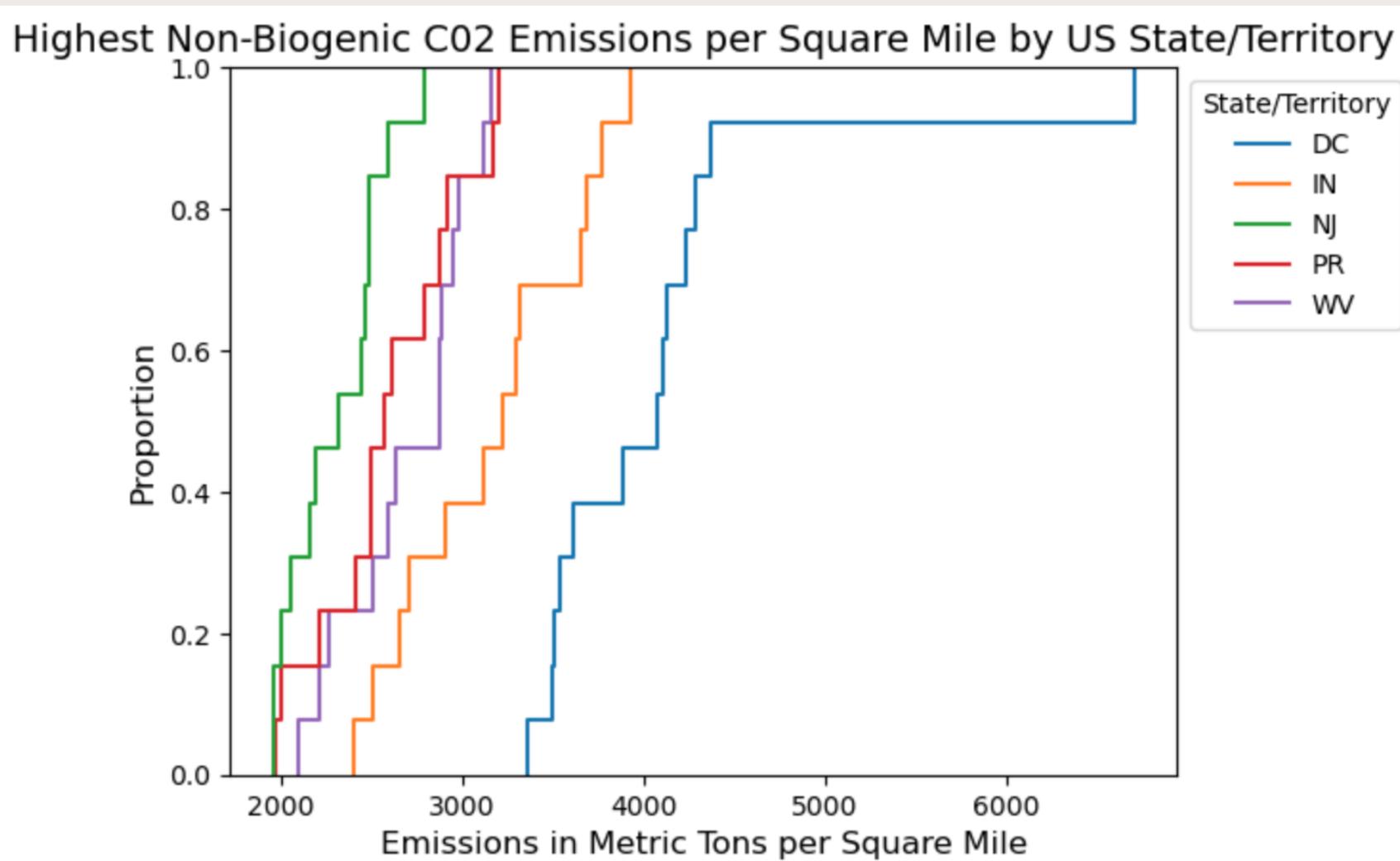


Question 3: ECDF, KDE with Log



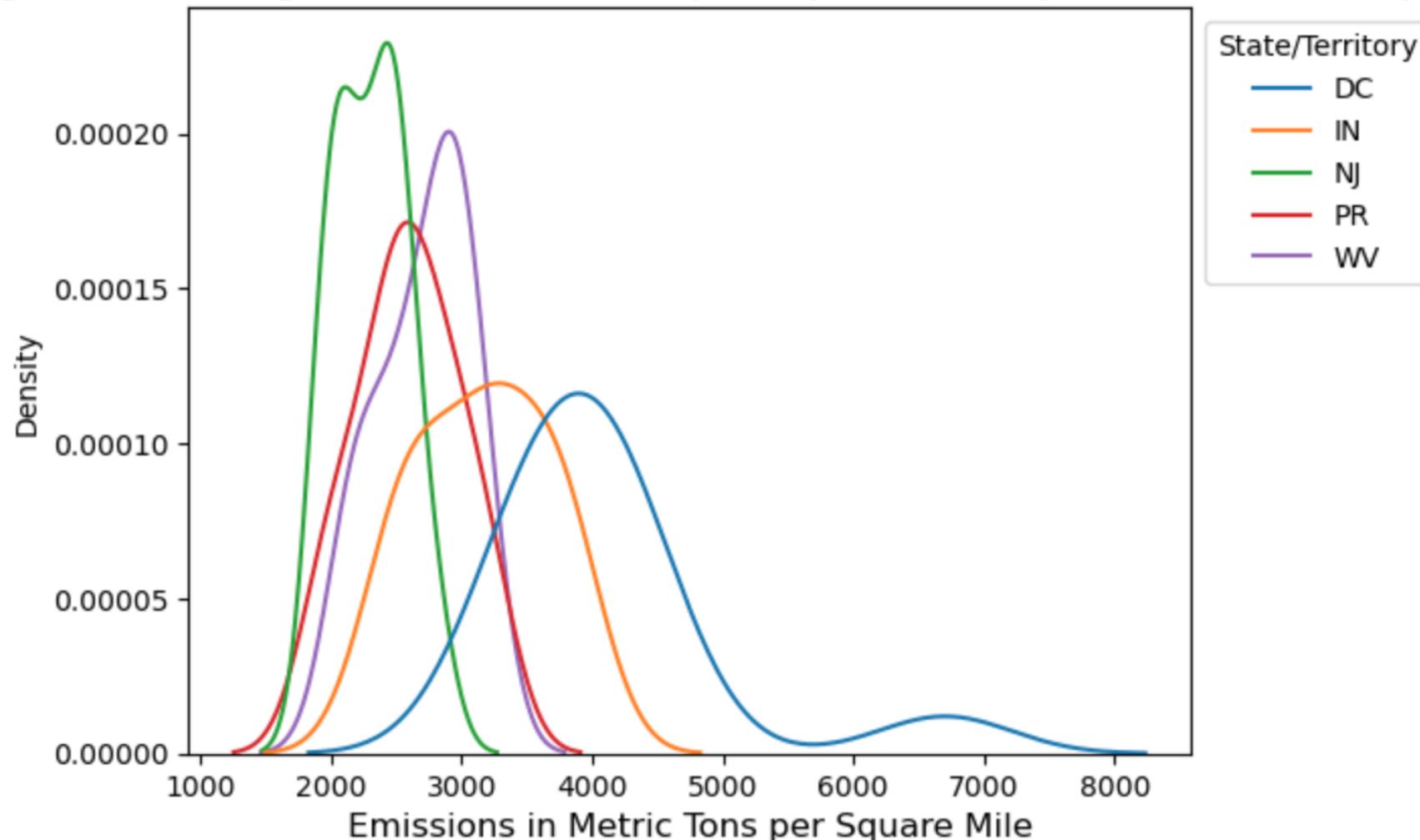
State/Territory
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
GU
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
PR
RI
SC
SD
TN
TX
UT
VA
VI
VT
WA
WI
WV
WY

Question 3: ECDF Normalized

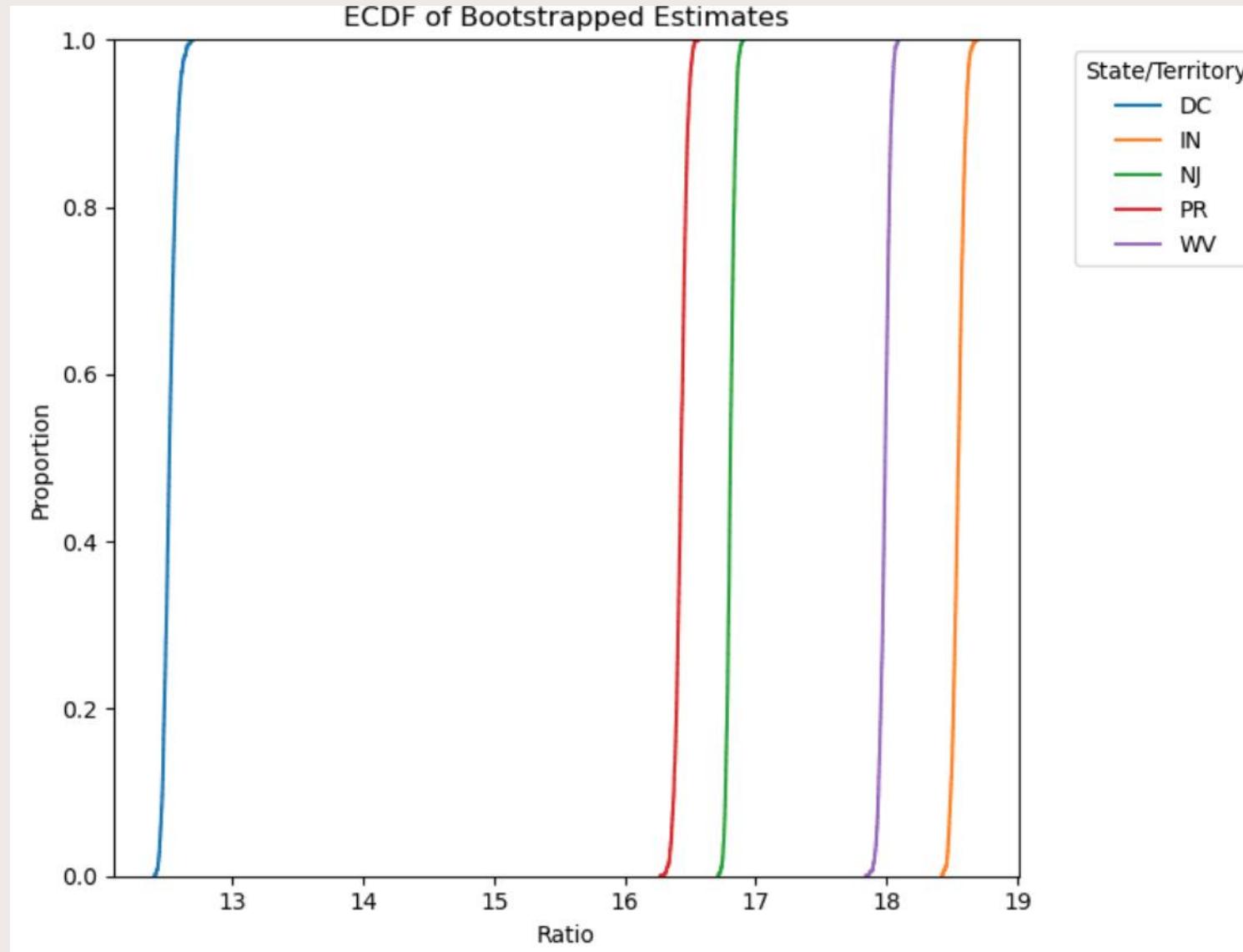


Question 3: KDE Normalized

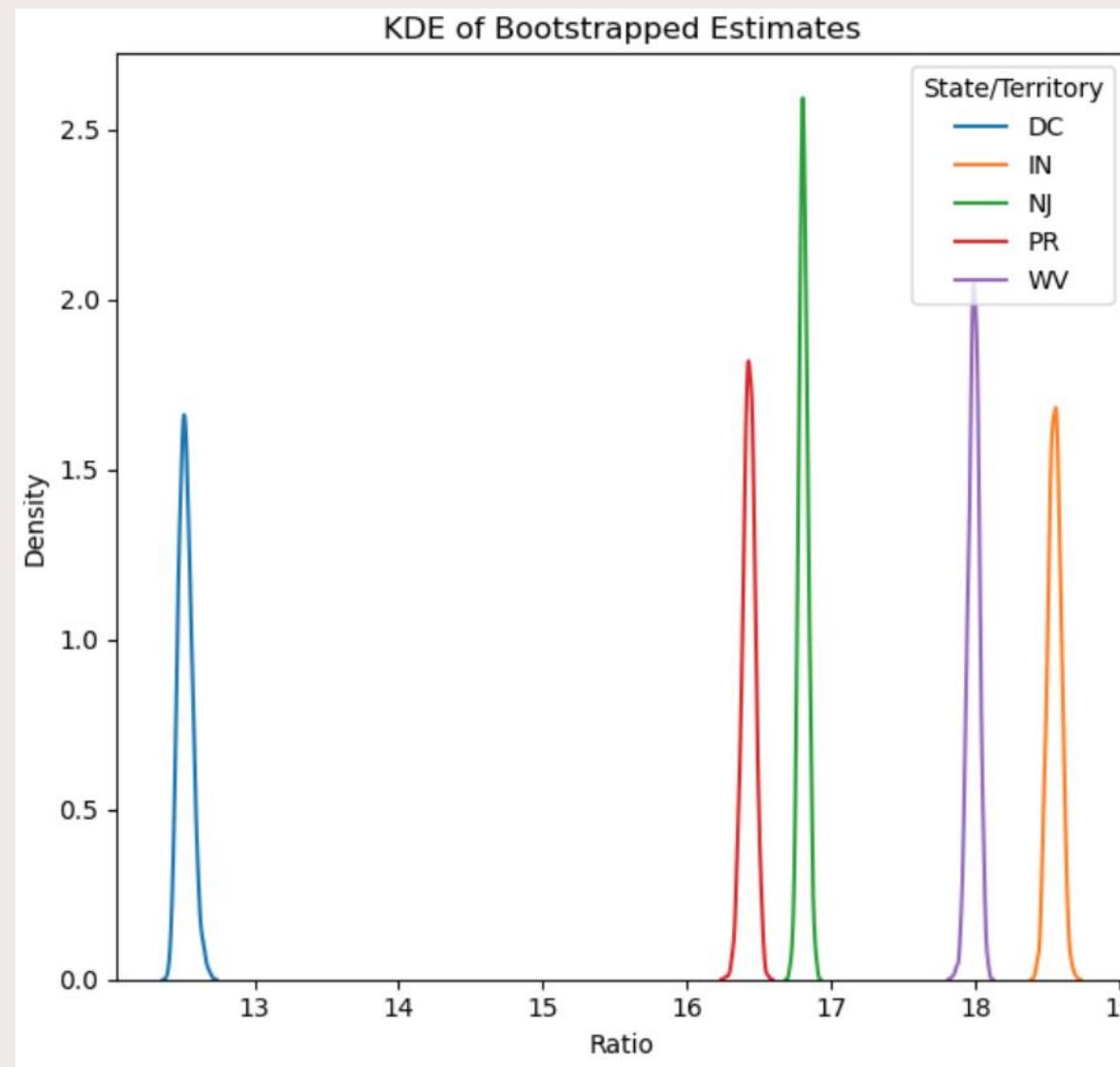
Highest Non-Biogenic CO₂ Emissions per Square Mile by US State/Territory



Question 4: Bootstrapped ECDF



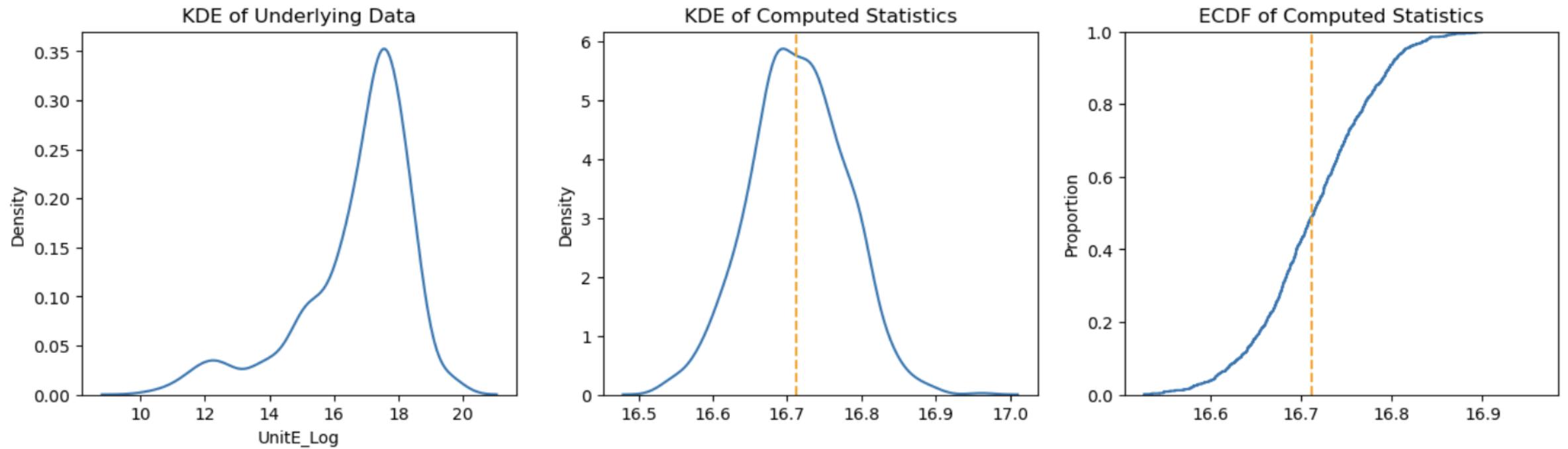
Question 4: Bootstrapped KDE



Question 5:

- Sequences do NOT have property of training data (when categorized by state) in bootstrap
- Small amount of entries per state, small sample size
- Bootstrapping minimally effective --> low variance per state
 - High KDE peaks
 - Smoother ECDF
- Therefore, estimates lack reliability and credibility
- Bootstrapping of all emissions data more effective

Question 4: Bootstrapping of overall data



Question 6:

Limitations

- Analysis was surface-level; hard to draw strong or actionable conclusions.
- Didn't incorporate outside factors like economic or policy data.

Future Work

- Combine emissions data with economic, regulatory, housing, and income variables.
- Explore how emissions change during economic or policy shifts.

Thank
you

